Fraud Detection in Credit Cards using Machine Learning

Seftico Frig Injek B
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
seftico.injek@binus.ac.id

Muhammad Fadzli Maula
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
muhammad.maula@binus.ac.id

Afdhal Kurniawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
afdhal.kurniawan@binus.ac.id

Karli Eka Setiawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
karli.setiawan@binus.ac.id

***Abstract*— Credit cards are the most popular payment method. With the rapid advancement of technology, the world is increasingly using credit cards instead of cash in daily life, which opens up numerous opportunities for fraudulent activities. Fraudulent transactions can take many different forms and fall under a number of different headings. This study employs several machine learning approaches. Three classification methods were utilized including Logistic Regression, XGBoost, and Gradient Boosting Classifier. The best classifier identified in this study is XGBoost. When comparing Accuracy, Precision, Recall, F1-score, MCC, and AUC, XGBoost demonstrates significantly higher values across all metrics.**

Keywords—Fraud Detection, Credit card, Logistic Regression, XGBoost, Gradient Boosting Classifier

I. INTRODUCTION

Financial fraud is a crime committed to obtain financial gain using deceptive and illegal methods. This fraud can be perpetrated in several financial domains, including business, banking, insurance, and taxation. Credit card fraud is a problem that has affected the payment system. Credit card fraud is a prevalent issue that leads to significant financial losses. The volume of online transactions has significantly increased, and online credit card transactions play a major role in this trend.

In recent years, the proliferation of digital transactions has transformed the global economy, but it has also led to a significant increase in credit card fraud, costing the global economy over $32 billion annually, with a projected rise to $38.5 billion by 2027[1]. In the United States alone, credit card fraud losses amounted to $11 billion in 2020[2]. Credit card fraud involves unauthorized transactions or using stolen card information, causing substantial financial losses and eroding consumer trust in digital payment systems.

Detecting and preventing such fraud requires advanced technological solutions, including machine learning algorithms, statistical techniques, and behavioral analysis, to analyze vast amounts of transaction data in real-time. To achieve this, we are using four methods of testing: Logistic Regression, XGBoost, and Gradient Boosting Classifier.

However, these systems face challenges with false positives, where legitimate transactions are incorrectly flagged as fraudulent, and false negatives, where actual fraudulent activities go undetected. For instance, a false positive rate of just 1% can affect millions of legitimate transactions given the volume of daily credit card usage[3]. Balancing the sensitivity and specificity of fraud detection algorithms is critical, as false positives can inconvenience customers and result in lost business for merchants, while false negatives lead to significant financial losses. This paper explores the methodologies employed in credit card fraud detection, examining their performance and addressing the issue of false positives and negatives, to enhance the accuracy and reliability of fraud detection mechanisms.

II. RELATED WORK

In recent years, there has been a lot of interest in using machine learning (ML) approaches for detecting financial fraud. Several studies have explored various machine learning models and their effectiveness in identifying fraudulent activities. For instance, a study by Hidayattullah et al. utilized meta-heuristic optimization in conjunction with Back Propagation Neural Networks (BPNN) and Support Vector Machines (SVM) to detect financial statement fraud. Their approach incorporated feature scaling and dimensionality reduction through Principal Component Analysis (PCA), which streamlined the classification model and significantly enhanced the performance of the fraud detection system. Notably, the optimized SVM classifier outperformed the BPNN, demonstrating substantial improvements in accuracy and Matthews Correlation Coefficient (MCC).

Bayesian optimization has also been employed to address unbalanced data in fraud detection. A study by Hashemi et al. proposed weight-tuning hyperparameters as a preprocessing step and recommended the use of CatBoost, XGBoost, and LightGBM for improved performance. Their evaluation on the "creditcard" dataset, comprising 284,807 transactions with 492 identified as fraudulent, demonstrated that leveraging LightGBM with class weight tuning improved fraud detection cases by 50% and the F1-score by 20% compared to recent methods. The combined use of LightGBM and XGBoost through majority voting achieved promising scores of 0.79 and 0.81, respectively, showcasing the effectiveness of ensemble learning approaches in enhancing fraud detection.

Systematic literature reviews (SLRs) have been instrumental in synthesizing existing research on financial

fraud detection using machine learning. Ashtiani and Raahemi conducted a thorough investigation into intelligent financial statement fraud detection (FSFD), focusing on the utilization of machine learning and data mining techniques. Their review of 187 papers highlighted a prevalent preference for structured data such as financial ratios over unstructured data. The study also emphasized the potential of ensemble methods, which leverage the collective power of multiple algorithms, for improved classification performance. However, the authors noted a significant gap in the utilization of unsupervised techniques like clustering, advocating for further exploration of these methodologies to unlock new insights into fraud detection.

Similarly, Ali et al. provided an exhaustive analysis of machine learning based methods for financial fraud detection, covering all major aspects of financial fraud activities. Their review synthesized 93 selected publications and identified essential issues, gaps, and limits in the field of financial fraud detection. The study underscored the adaptability of various machine learning techniques across different types of financial fraud, including credit card fraud, health insurance fraud, and cyber financial fraud. Notably, Support Vector Machine (SVM) and Artificial Neural Network (ANN) emerged as prevalent techniques across multiple fraud types, highlighting their robustness and effectiveness in fraud detection scenarios.

These comprehensive analyses and reviews have provided valuable insights and guidance for future research endeavors in the field of financial fraud detection. They underscore the importance of employing advanced ML methodologies and exploring diverse datasets to enhance the accuracy and efficacy of fraud detection systems.

III. Methodology

The technique proposed in this paper is used to detect fraud in credit card systems. This paper compares various machine learning algorithm such as XGBoost, Logistic Regression and Gradient Boosting Classifier to determine which algorithm are most suitable and can be adapted by credit card merchants in identifying fraudulent transactions. Figure 1 shows an architectural diagram that depicts the overall system framework.

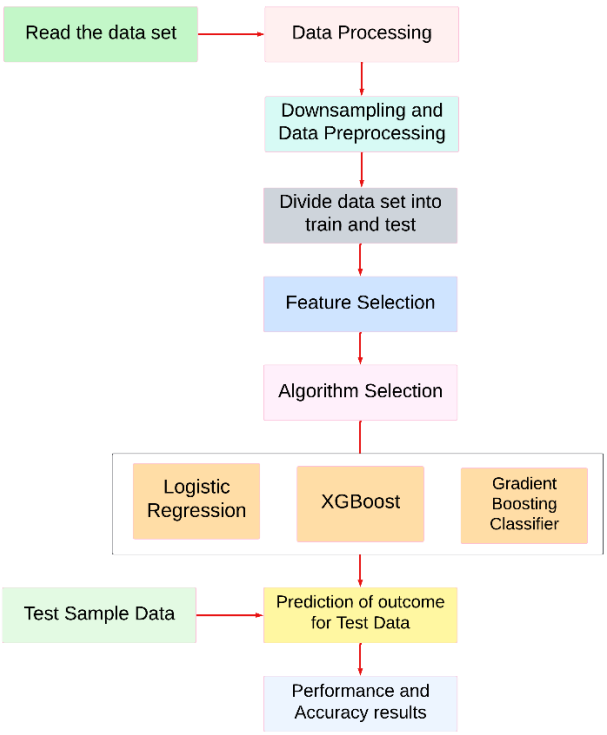

Fig. 1. System Architecture

A. *Data Description*

In this research the Credit Card Transactions dataset was used, which can be downloaded from Kaggle [14]. This dataset contains almost 20 million transactions resulting from IBM's multi-agent virtual world simulation.

The dataset was created combining three data sources; users, card and credit card information with the transaction record data. The credit card transaction dataset contains 20 million records of credit card transactions. The Users dataset contains information about the simulated people, including gender, age, income, etc. The Card dataset containing information about the simulated credit cards.

B. *Data Preparation*

The three collected dataset will undergo data cleaning first. Below are the steps taken for each dataset:

Data Preprocessing for Users Dataset (User Information) - Create a user ID 'User' to be used as the primary key to join the Users table with the Card table. Next, remove the dollar sign before 'Yearly Income – Person', 'Total Debt', and 'Per Capita Income – Zipcode'. Additionally, new columns are created, such as Retired, Person Location Income ratio, Person Income to Debt, and Location Income to Debt. Retired is set to 'Yes' if the current age of the user is greater than the retirement age. The Person Location Income ratio is calculated using the formula (1).

$$\frac{(Yearly\ Income-Person)}{(User\ Location+0.01)} \quad (1)$$

The Person Income to Debt ratio is calculated using the formula (2).

$$\frac{(Yearly\ Income-Person)}{(Total\ Debt+0.01)} \quad (2)$$

The Location Income to Debt ratio is calculated using the formula (3).

$$\frac{(User\ Location\ Income)}{(Total\ Debt+0.01)} \quad (3)$$

These variables are created because they can potentially represent the income level and financial pressure of users, which can be associated with fraudulent transactions.

Data Preprocessing for Card Dataset (Card Information) - Create a user index 'User_Card'. This index will be connected to the Users table using the ID created in the previous step. After joining with User Information, the same actions are performed here as in the previous step, such as removing the dollar sign from the Credit Limit column. Next, select the variables of interest, namely Card Brand, Card Type, and Credit Limit in the Cards table.

Data Preprocessing for Credit Card Transaction Dataset - Join the table with Card and User Information. Remove the dollar sign from the Amount column and remove data with negative transaction amounts. Next, remove columns such as errors and merchant names; these columns either have too many missing values or are hard to interpret. After that, create the day of the week and create the time of day (when the transaction occurs, separating a day into 8 periods): Midnight (23-2), Early Morning (2-5), Morning (5-8), Late Morning (8-11), Noon (11-14), Afternoon (14-17), Evening (17-20), Late Night (20-23). Create the last digit of the transaction amount because the last digit of the transaction amount can be a trait of fraud. The next step is to remove columns that will not be used, such as Time and Transaction_Time.

The above steps include processes such as data cleaning and data combination. These steps are the initial steps to make the fraud detection model more efficient and accurate. During data cleaning, records with transaction amounts < 0 are removed, as well as removing dollar signs among other things. This is done to ensure the data is consistent and to improve the accuracy of the model's predictions. Next, data combination involves merging files containing user and credit card information with the transaction record data.

One of the most difficult aspects of fraud detection is the class imbalance, which occurs when fraudulent transactions account for a small percentage of total transactions. To address this, we employ Downsampling, a technique used to reduce the size of a dataset by decreasing the number of samples while maintaining a proper representation of the minority class, in this case, fraudulent transactions. Downsampling is performed to balance the dataset at 500,000 records, with 5% being fraudulent transactions. This means that out of 500,000 records, 25,000 are fraudulent transactions, and the remaining 475,000 are non-fraudulent transactions.

Transforming the data involves using one-hot encoding for categorical variables and scaling for numerical variables. One-hot encoding converts categorical variables into binary numerical representations, while scaling transforms numerical variables to be within the same scale, either through normalization (range [0,1]) or standardization (mean of zero and standard deviation of one).

The dataset is then split into training and testing sets, with 80% of the dataset used for training the model and 20% for testing the model's performance. The split maintains the class distribution, ensuring that both sets have approximately 5% fraudulent transactions. This balance is crucial for training the model to recognize patterns of fraudulent behavior and testing its performance on similarly distributed data.

C. Modelling and testing

These models were chosen for their distinct characteristics including Logistic Regression,XGBoost and Gradient Boosting Classifier. Each model was trained and tested on a consistently preprocessed dataset to ensure a fair comparison of their performance.

XGBoost is a powerful boosting algorithm known for its efficiency and performance on large datasets. The implementation involved initializing the XGBoost classifier, training it on the training data, making predictions on the test data, and evaluating its performance using various metrics like accuracy, precision, recall, F1-score, and the confusion matrix.

Logistic Regression is a simple yet effective statistical method for binary classification. Following the implementation of XGBoost, the logistic regression model was trained on the same dataset and evaluated using similar performance metrics. Despite its simplicity, logistic regression provides a solid baseline for comparing the performance of more complex models.

Gradient-Boosting Classifier This ensemble technique generates models in a sequential manner, with each new model seeking to fix the flaws of the preceding ones. The Gradient Boosting Classifier enhances model accuracy while reducing overfitting by integrating the predictive capacity of numerous weak learners, most commonly decision trees. It was trained using training data, as with prior models, and evaluated using accuracy, precision, recall, F1-score, and the confusion matrix.

Model Evaluation

Each model's performance was evaluated using the following metrics:

- Accuracy: The ratio of correctly identified cases to the total number of cases.
- Precision: The ratio of true positive cases to all cases predicted as positive. High precision means fewer false positives.
- Recall: The ratio of true positive cases to all actual positive cases. High recall means fewer false negatives.
- F1-score: The harmonic mean of precision and recall, balancing the two metrics. This is particularly useful for imbalanced class distributions, providing a single performance measure. The F1-score formula (4).

$$\text{F1-score} = 2 \text{ x } \frac{Precision\ x\ Recall}{Precisian+Recall} \quad (4)$$

- Matthew's correlation coefficient (MCC): A metric that considers true and false positives and negatives. It is viewed as a balanced measure, even when class sizes are highly imbalanced. The MCC ranges from -1 to +1, where +1 signifies perfect prediction, 0 signifies random prediction, and -1 signifies

complete disagreement between predictions and actual outcomes.

- ROC AUC (Receiver Operating Characteristic Area Under Curve): This metric evaluates the model's ability to distinguish between classes. It measures the separability achieved by the model. A higher AUC indicates better performance in correctly predicting positive instances as positive and negative instances as negative. An AUC of 1 represents a perfect model, while an AUC of 0.5 indicates performance equivalent to random guessing.

Training and Testing Split. To evaluate model performance on previously unseen data, the dataset was divided into training and testing sets at an 80:20 ratio. This split ensures that the model is tested on data that it did not encounter during training, resulting in an unbiased assessment of its performance.

IV. Result

We employed downsampling and data transformation techniques, including one-hot encoding and scaling, to enhance data quality and machine learning model performance. The metrics used to assess the models include Accuracy, Precision, Recall, F1 Score, MCC (Matthews Correlation Coefficient), and AUC (Area Under the ROC Curve). Each metric provides insight into various aspects of the model's performance, which is crucial in the context of fraud detection where the cost of false negatives (missed frauds) can be very high. The evaluation results are presented in Table 1.

TABLE I. COMPARISON OF MODEL

	Accuracy	Precision	Recall	F1 Score	MCC	AUC
Logistic Regression	0.96281	0.818498	0.3292	0.469548	0.505142	0.942057
XGBoost	**0.97837**	**0.865876**	**0.6714**	**0.756337**	**0.751816**	**0.984313**
GradientBoostingClassifier	0.96991	0.826500	0.5040	0.626165	0.631826	0.957056

Accuracy measures the amount of correctly classified instances among the total instances. While XGBoost has the highest accuracy, it's important to note that accuracy alone may not be sufficient for imbalanced datasets, such as fraud detection, where the number of non-fraudulent transactions significantly outweighs fraudulent ones. As illustrated in Fig. 2, the accuracy comparison of the three models reveals no significant differences. However, XGBoost emerges as the top performer with an accuracy of 0.9784.

Precision indicates the proportion of true positive estimates out of all positive predictions.. Higher precision is crucial in fraud detection to reduce the number of false positives. Fig. 3 showcases the model evaluation metrics using a heatmap. Logistic Regression and Gradient Boosting Classifier have slightly lower precision values and XGBoost again leads with a precision of 0.8659. XGBoost achieves the highest precision, suggesting it is more effective at correctly identifying fraudulent transactions without falsely labeling non-fraudulent ones as fraudulent.

Comparing the Precision, recall, F1-score, MCC, and AUC, it can be seen in Fig. 3 that XGBoost has better values compared to the other two models. XGBoost consistently performs better than the other models across all metrics. The Gradient Boosting Classifier also performs well but falls short of XGBoost in most metrics. Furthermore, Fig. 4 displays a bar chart of all metrics. Comparing the three models, XGBoost has a much higher bar chart for all the metrics. The bar chart clearly demonstrates that XGBoost outperforms both Logistic Regression and the Gradient Boosting Classifier across all evaluation metrics. This makes it the most effective model for fraud detection in this context.

V. Conclusion

In this paper, we studied the credit card fraud detection problem. We introduced a machine learning method to enhance the effectiveness of fraud detection. We used the credit card dataset obtained from Kaggle. This data contains over 20 million transactions generated from a multi-agent virtual world conducted by IBM. We proposed three methods for fraud detection.

We evaluated these models using several performance metrics including Accuracy, Precision, Recall, F1-score, MCC, and AUC. The results demonstrated that XGBoost outperformed both Logistic Regression and Gradient Boosting Classifier across all evaluation metrics. Specifically, XGBoost achieved the highest all evaluation metrics, indicating its superior ability to accurately identify fraudulent transactions while minimizing false positives and negatives. The accuracy for Logistic Regression, XGBoost, and Gradient Boosting Classifier are 0.9628, 0.9784, and 0.9699 respectively. These findings highlight the importance of using advanced ensemble method like XGBoost for fraud detection in highly imbalanced datasets. Performance of XGBoost in all key metrics makes it the most suitable model for this task, offering a reliable and robust solution for detecting fraudulent transactions in large-scale credit card data. This study can be enhanced in the future by adding more classifiers. This addition in classifiers is likely to reveal additional insights about the result.

References

[1] Credit card fraud costs global economy $32bn: AMF. (2023b, October 10). HiDubai Focus. https://focus.hidubai.com/credit-card-fraud-costs-global-economy-32bn-amf/

[2] E. Morgan, R. (2021). Financial Fraud in the United States (NCJ 255817). U.S. Department of Justice. Retrieved April 28, 2024, from https://bjs.ojp.gov/content/pub/pdf/ffus17.pdf

[3] Wedge, R., Max Kanter, J., Veeramachaneni, K., Moral Rubio, S., & Sergio Iglesias Perez, S. (2017). Solving the "false positives" problem in fraud prediction. Automated Data Science at an

Industrial Scale, ArXivID 1710.07709. http://arxiv.org/abs/1710.07709

[4] S. Hidayattullah, I. Surjandari, and E. Laoh, "Financial statement fraud detection in indonesia listed companies using machine learning based on meta-heuristic optimization," in 2020 International Workshop on Big Data and Information Security, IWBIS 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 79–84. doi: 10.1109/IWBIS50925.2020.9255563.

[5] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud Detection in Banking Data by Machine Learning Techniques," IEEE Access, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.

[6] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review," IEEE Access, vol. 10. Institute of Electrical and Electronics Engineers Inc., pp. 72504–72525, 2022. doi: 10.1109/ACCESS.2021.3096799.

[7] A. Ali et al., "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," Applied Sciences (Switzerland), vol. 12, no. 19. MDPI, Oct. 01, 2022. doi: 10.3390/app12199637.

[8] D. Trisanto, N. Rismawati, M. F. Mulya, and F. I. Kurniadi, "Effectiveness Undersampling Method and Feature Reduction in Credit Card Fraud Detection," International Journal of Intelligent Engineering and Systems, vol. 13, pp. 173–181, 2020, doi: 10.22266/ijies2020.0430.17.

[9] A. Priyadi, I. A. Hanifah, and M. Muchlish, "The Effect of Whistleblowing System Toward Fraud Detection with Forensic Audit and Investigative Audit as Mediating Variable," Devotion : Journal of Research and Community Service, vol. 3, no. 4, pp. 336–346, Feb. 2022, doi: 10.36418/DEV.V3I4.121.

[10] I. A. Yuri and M. R. Sari, "Fraud Awareness and Fraud Detection-Prevention Methods in The Indonesian Private and Public Sector," Global Financial Accounting Journal, vol. 6, no. 1, pp. 100–107, Apr. 2022, doi: 10.37253/GFA.V6I1.6529.

[11] B. An and Y. Suh, "Identifying financial statement fraud with decision rules obtained from Modified Random Forest," Data Technologies and Applications, vol. 54, no. 2, pp. 235–255, Jun. 2020, doi: 10.1108/DTA-11-2019-0208.

[12] M. E. Lokanan and K. Sharma, "Fraud prediction using machine learning: The case of investment advisors in Canada," Machine Learning with Applications, vol. 8, p. 100269, Jun. 2022, doi: 10.1016/J.MLWA.2022.100269.

[13] A. M. Khedr, M. El Bannany, S. Kanakkayil, and M. El Bannany, "PREPRINT An Ensemble Model for Financial Statement Fraud Detection An Ensemble Model for Financial Statement Fraud Detection", doi: 10.3897/arphapreprints.e69590.

[14] "Credit Card Transactions." [Online]. Available: https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions

[15] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud Detection in Banking Data by Machine Learning Techniques," *IEEE Access*, vol. 11, pp. 3034–3043, Jan. 2023, doi: 10.1109/access.2022.3232287.